

Particle launches on token efficiency as frontier models get pricier per task and AI security budgets grow

VC Tech Radar

2026-10-29

Particle launches on token efficiency as frontier models get pricier per task and AI security budgets grow

By VC Tech Radar • October 29, 2026

Suhail's Particle launched with a model that cuts token use on top of DeepSeek, new benchmarks show long-horizon coding agents still struggle, and a16z's CISO survey shows where AI-security budgets are going. Also covered: new rounds in robotics, vertical agents and defense.

Particle bets on cheaper intelligence per task

Suhail launched a new company, Particle. Its first model, particle-1, is the start of a developer-focused series. The company says it matches performance while using about 45% fewer tokens [1]. It is post-trained on DeepSeek V4.1 Flash 1017. The stated long-term goal is "continuous post-training," meaning model improvements that ship as fast as software [2]. OpenRouter relays Particle's claim of 28–52% fewer output tokens than the DeepSeek base across four coding benchmarks [3]. Garry Tan says Particle will be at the frontier for the lowest dollar per unit of intelligence [4].

The underlying problem is cost per task, not price per token. Two Minute Papers says the new DeepSeek 4.1 Flash 1017 builds much better working software and costs just over \$1 per million output tokens. In the host's own tests, though, it took two to three times as many tokens as Claude Opus to finish the same task [5]. Particle's pitch targets exactly that gap: open weights with better token efficiency.

Other model signals from the period: - **Gemini 4 Carbon**. Demis Hassabis calls it "SOTA across key benchmarks" but published no figures [6]. - **ViBench**

2.0. Replit’s new benchmark has agents build a whole product from an empty project, one feature at a time. Even GPT-6.1 Sol, GPT-6 Astra and Opus 5.5 passed every check in only 49–65% of builds [7]. Long-horizon autonomous building still has a lot of room to improve. - **OpenAI and Google TPUs.** Reuters, citing The Information, reports that OpenAI plans to buy Google TPUs. No terms were disclosed and neither company commented [8]. - **Cartesia H-Nets.** Albert Gu reports that a new version of these byte-level models shows up to 3–5x efficiency gains in controlled scaling-law experiments. At larger scale it matches an open-source model trained on 11T tokens using about 10x less data and compute [9].

AI security: budgets are rising, but incumbents get the first call

a16z surveyed 100 CISOs at companies with 1,000+ employees. 87% expect security spending to rise, driven mainly by AI-enabled threats and by their own AI deployments [10]. All seven of the most-selected 2027 initiatives involve AI [10]. Production use is still low: 18% run AI SOC tools in production, and 13% run AI offensive security or AI vulnerability remediation. Within 12 months, 45% plan to buy or upgrade internal AI-agent security, 42% offensive security, and 38% AI governance [10].

The go-to-market picture is less favorable for startups: - 44% would first look for AI security features from their existing platform; 21% would start with an AI-native vendor. 47% say a capability gap their incumbent can’t fill would lead them to pick a startup [10]. - Some categories have as many as 43 vendors [10]. - a16z discloses that it led Armadin’s Series B. Armadin uses autonomous agents to find exploitable paths through enterprise systems [10].

Attack capability is the reason for the urgency. In a simulated 32-step network attack, GPT-4o stalled at step 2. Today’s top models got about three-quarters of the way through on average, and the best runs completed all 32 steps [11]. Wiz’s CTO said the company’s Atlas harness, which combines DeepMind, OpenAI and Anthropic models, made its PHP remote-code-execution finding possible. He described the flaw as affecting 12 years of PHP deployments in default configurations [12]. He said most monitored workloads were still unpatched [12]. He told startups to plan for critical vulnerabilities arriving monthly or even weekly, and to build automated patching and infrastructure-as-code while they are still small [12].

Rounds and launches

- **Foundry Robotics** raised a \$100M Series A led by MarlinSPIKE, with Khosla continuing. It is building a 130,000 sq ft factory in Emeryville and says it already ships to paying customers in energy, telecom, autonomous vehicles and defense [13].

- **Broccoli (YC W22)** raised \$52M across its Series A and B for AI phone agents that serve plumbing, HVAC and electrical contractors. It has 1,500+ businesses on the platform and has booked 2.5M appointments. The founders knocked on 140 contractor doors, and the plumbing owner who first gave them time is now COO [14].
- **Fragment** launched with \$56M: a Series A led by 8VC, plus a seed coled by General Catalyst and A* [15]. It resolves supply-chain exceptions, starting with direct procurement [16].
- **Formation Bio** launched a \$600M fund to acquire drugs and move them through the clinic, pairing the capital with its AI platform [17].
- **Diode**, a PCB design-and-manufacturing company whose Series A a16z led, hired an a16z investor as president [18].
- **Bunkerhill**, which builds agents for health systems, says it has grown 12x annualized since January and closed eight figures of ARR in one week [19].
- **FluxCap (YC P26)**, founded by former DOJ antitrust chiefs, came out of stealth building “Deep Legal Tech” [20].

Operating data from scaled vertical and defense companies

Hippocratic AI has about 300 employees, has raised about \$450M, and launched its product only about 21 months ago. It reports 290M patient interactions and counts five of the top seven payers among its customers [21]. The CEO says it charges about \$9–10/hour. Running the same 31-model stack on closed models would cost about \$105/hour in compute alone, which is why it uses open weights [21]. Most of its traction comes from “infinite pilot” work, meaning care nobody offered before. In one example, calls to 1,700 patients lost to follow-up led 180 to complete CT scans, which found three cancers [21].

Stark, a European defense company backed by Sequoia, won billions in contracts within 18 months [22]. It grew from zero to 700 people in two years [22] and plans to invest about €600M in R&D and production capacity over 12–36 months [22]. The founders say European procurement has gone from years to months, and they expect that change to last [22].

Pressure on incumbent software

SaaStr, citing the WSJ, says about \$300B in risky PE-backed loans come due in 2028–29, and roughly a third of that is software [23]. Software loans fell from about 95 to about 85 cents on the dollar after new Anthropic and OpenAI models came out, while other loans held near 96 [23]. Elizabeth Yin argues customers will no longer pay \$25,000 for a CRM they can rebuild themselves [24]. She says Hustle Fund runs about 90% of its operations on agents [25]. In her view the lasting value moves to support: maintenance, security and uptime [26].

Orbital compute

Cowboy Space, led by Robinhood cofounder Baiju Bhatt and backed by Index, unveiled a rocket built to carry five racks of NVIDIA Vera Rubin GPUs into orbit. Its stated aim is space compute that costs the same as compute on the ground [27]. Starcloud, a competitor backed by Benchmark, is valued at \$2.3B [27].

Sources

1. X post by @Suhail
2. X post by @Suhail
3. X post by @OpenRouter
4. X post by @garrytan
5. Deepseek’s New AI Has A Secret
6. X post by @demishassabis
7. X post by @Replit
8. r/artificial post by u/winer666
9. X post by @_albertgu
10. X article by @steph_zhang
11. X post by @a16z
12. Wiz’s CTO on PHP2Shell and the coming ‘vulnpocalypse’ | Equity Podcast
13. X post by @adarshkulkarni
14. X post by @ycombinator
15. X post by @pranavmulgund
16. X post by @roelofbotha
17. X post by @a16z
18. X article by @zabie_e
19. X post by @nish_khandwala
20. X post by @ycombinator
21. Why vertical AI apps will win | Munjal Shah, Co-founder and CEO, Hippocratic AI
22. European Defense Sovereignty Starts in the Factory: Stark’s Uwe Horstmann and Johannes Schaback
23. The \$300 Billion Software Debt Wall: The Big B2B Defaults Unfortunately Haven’t Even Started Yet
24. X post by @dunkhippo33
25. X post by @dunkhippo33
26. X post by @dunkhippo33
27. The New Space Race: Startups Press Ahead on ‘Orbital Infrastructure,’ Test Feasibility of Data Centers